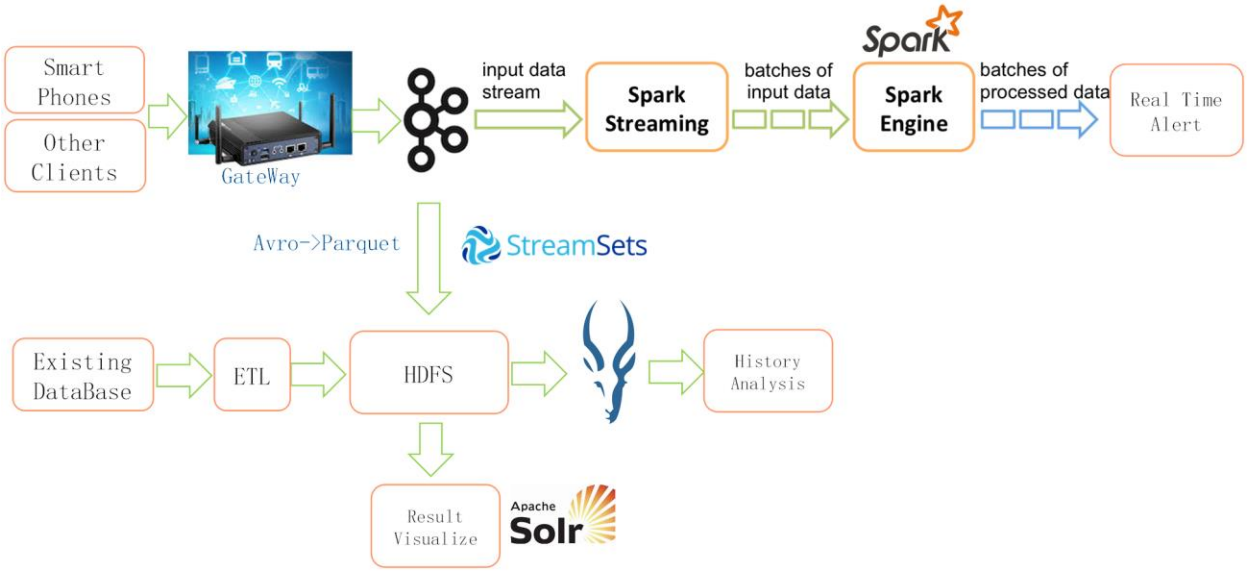
一种可用于 2019 武汉冠状病毒疫情控制的大数据处理平台参考设计

项目目的
本项目旨在提议一个可用于抗击 2019 武汉冠状病毒的大数据处理平台，用于处理全民疫情大数据，记录全民的疫情相关数据，并处理这些数据：

1. 实时挖掘可能的疫情传染警报信息通知相关人员。
2. 分析感染人员的历史数据，生成疫情接触感染树，用于指导发现疑似病例及时采取隔离措施以降低疫情进一步传染扩散。

本项目处在早期构思阶段，目前仅有一些前期类似项目所用的源代码可参考或重用于本平台的开发。
数据模型所涉及数据所有单位较多，数据获取比较困难，本设计只是提出一个能够处理这些数据的系统架构，具体的数据获取途径与权限不在本项目的设计范畴。
欢迎感兴趣的朋友一起参与开发，也希望本项目能对正在进行类似项目开发的相关研发单位的开发人员有一定的参考作用。

大数据处理平台系统架构参考设计



系统工作流程：
数据导人与实时流处理阶段：
实时动态获取的原始信息数据源通过网关（Gateway）上传到数据中心的 Kafka 服务器收集并统一管理与分发，原始数据可以采用 avro 格式编码。

新导入数据首先在流处理引擎（这里为 spark streaming）里被用于实时分析，如果有危险情况立即发送信息给相关人员：
个体快速实时分类（初步分析）
如果自己是病源给当前在同一个范围内的人发送信息已警示危险，同时抄送公安机关。

如果自己所在区域有其他病例，则给本人发送信息告知危险，同时抄送公安机关。

Kafka 里的数据除了用于实时分析还被通过 streamSets 转换成 Parquet 格式并导入到大数据分布式存储系统 HDFS，用于后续历史数据批处理(这里使用 Impala 用于数据表格查询)并生成感染树：
深入挖掘数据信息，以及关联信息。
生长病例关联树：接触记录，扩散范围 =》 生成数状图，人员关系视图，地图视图

除了通过 kafka 服务器收集实时数据，本系统平台还支持将已有数据库管理平台的数据 ETL 导入到本平台中，并统一交由批处理服务器（这里为 Impala）进型历史数据处理。

最后生成的人员关系图以及地图视图的可视化工作在 Solr 服务器中完成。

数据模型 schema 总览

姓名	手机号	蜂窝漫游记录	电子支付交易记录	公共交通出行记录	个人交通出行记录（GPS 地图记录）	高危险地出现记录	电商购物交易记录	诊疗记录	户籍地址	近期居住地址	亲密接触者	身份证号码
		是否漫游| 漫游地| 时间	支付种类 | 支付时间| 支付双方地址	车次或车牌号 | 上车时间戳 |下车时间戳	车牌号 | Location | time stamp | GPS 数据存储地址	地点类型 | 地点| 到达时间 |离开时间	交易商品名称| 交易发生地|收货地址| 收货电话| 收货人|交易时间| 快递到达时间	状态（正常，亲密接触者疑似，确诊轻症，确诊重症，治愈出院，死亡）| 体温 | 测量地点 | 测量时间				

流数据 schema 架构与算法（用于判断实时数据的状态，并在必要情况下生成实时警报信息）

姓名	手机号	诊疗状态记录（0. 正常，1.亲密接触者疑似，2.确诊轻症，3.确诊重症，4.治愈出院，5.死亡）	最近一次体温测量值	体温测量地点	体温测量时间	户籍地址	近期居住地址	亲密接触者	身份证号码	记录上传时间
String	String	Char	float	String	TS	String	String	String	String	TS

基于分析诊疗状态记录的简单快速处理算法：

1. 如果不为空而且!=0 则发出警报信息给相关人员包含：身份证，当前地点（查询位置表格），诊疗状态
2. 如果诊疗状态为空，则判断体温，如果体温>37.5 摄氏度则发出警报信息给相关人员：身份证，当前地点（查询位置表格），体温值

历史数据架构与算法

公共交通出行记录数据 schema

姓名	手机号	交通统计标识（车牌号，航班号，或者线路号）	上车时间	下车时间	记录上传时间	身份证号码
String	String	String	TS	TS	TS	String

个人出行记录数据 schema（HBase 可以用于存储 GPS 轨迹数据）

姓名	手机号	交通统计标识（车牌号，共享单车，共享电瓶车）	开始时间	结束时间	记录上传时间	身份证号码	GPS 轨迹
String	String	String	TS	TS	TS	String	HEX

高位地点出现记录数据 schema

姓名	手机号	高位地标识（0 超市，1 医院，2 车站）	开始时间	结束时间	记录上传时间	身份证号码	GPS 轨迹
String	String	Char	TS	TS	TS	String	HEX

电商购物交易记录数据 schema

姓名	手机号	交易商品名称	交易发生地	收货地址	收货电话	收货人	交易时间	快递到达时间	记录上传时间	身份证号码
String	String	String	String	String	String	String	TS	TS	TS	String

电子支付交易记录数据 schema

姓名	手机号	支付种类(0 微信，1 支付宝，其它)	支付方地址	收款方地址	交易时间	快递到达时间	记录上传时间	身份证号码
String	String	Char	String	String	TS	TS	TS	String

蜂窝漫游数据表格

姓名	手机号	是否漫游	漫游区域	漫游状态变更发生时间	记录上传时间	身份证号码
String	String	bool	String	TS	TS	String

历史数据的流行病史分析算法（初步想法，待完善，需要医疗相关专业人士帮助）：
如果手机漫游到过武汉，或者有 GPS 至武汉的记录，或者现场交易电子支付记录，检查行踪范围是否为被污染区，或者是否有跟病例交易记录。
公共交通记录航班号或者车次是否有病例同乘。